Day 2: Research Design Issues in Data Science

Kenneth Benoit and Slava Mikhaylov

Introduction to Data Science and Big Data Analytics

18 August 2015

Day 2 Outline

Statistical learning and data science

Research Design

Benchmark of randomized trials

- The Selection Problem
- Selection bias and random assignment
- Observational studies
- A good research project

Review of basic of statistical theory

Introduction to R and RMarkdown

Statistical learning and data science

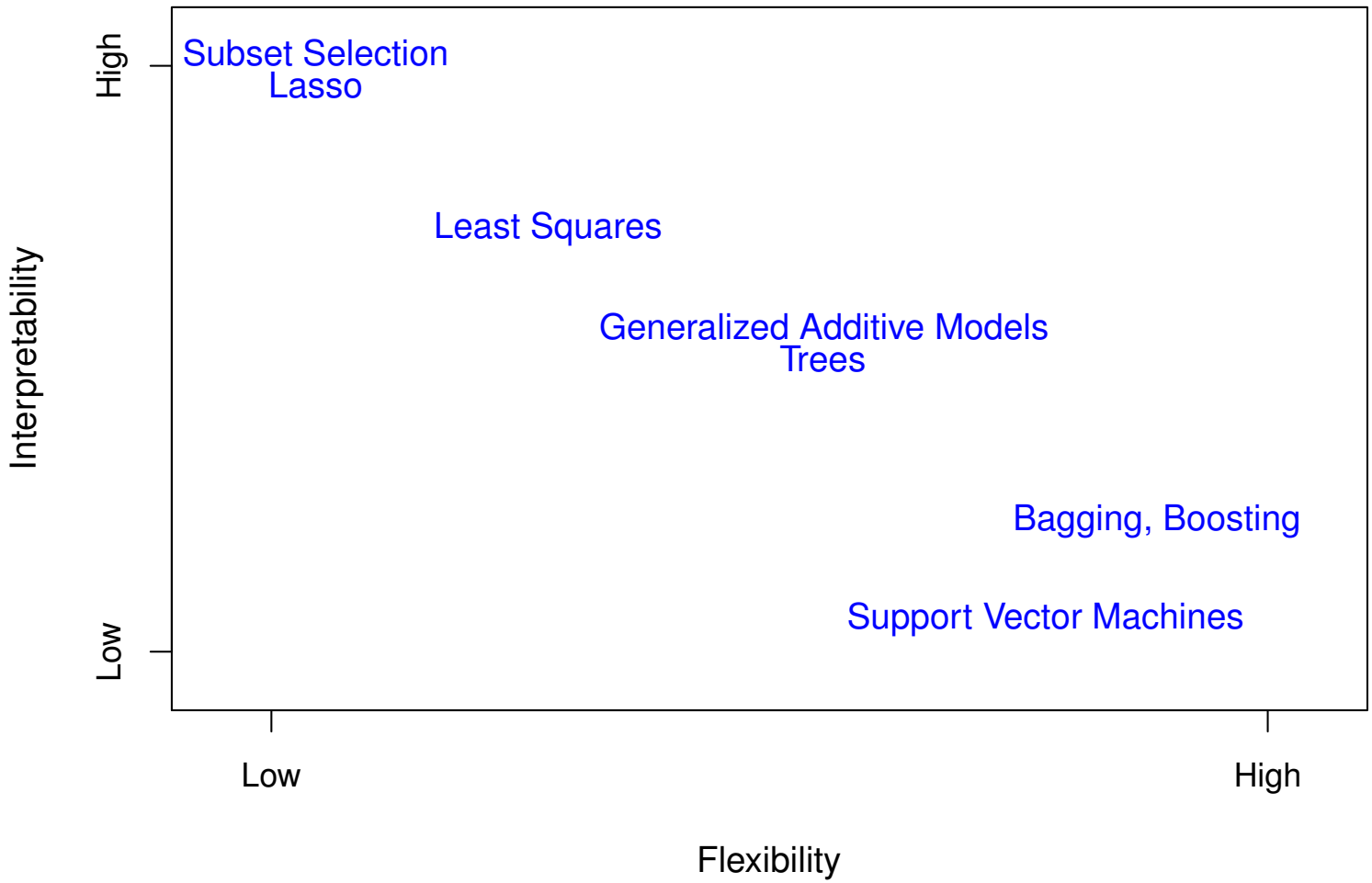
Statistical learning

- Statistical learning refers to a vast set of tools for *understanding data*.
- For a quantitative response Y and a set of predictors X:

$$Y = f(X) + \epsilon$$

- Here, f represents the *systematic* information that X provides about Y.
- Statistical learning refers to a set of approaches for estimating f.
- Most of the next two weeks we'll spend talking about different ways to estimate f and how to evaluate whether we've done a good job with it.

Statistical learning trade-offs



Where does this course fit in?

- Supervised versus unsupervised learning.
- Regression versus classification.
- No single *best* method. We'll spend a lot of time choosing the most appropriate tool for a given dataset using different measures of the quality of fit. E.g. MSE

$$MSE_{training} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

$$MSE_{test} = \mathrm{Ave}(\hat{f}(x_0) - y_0)^2$$

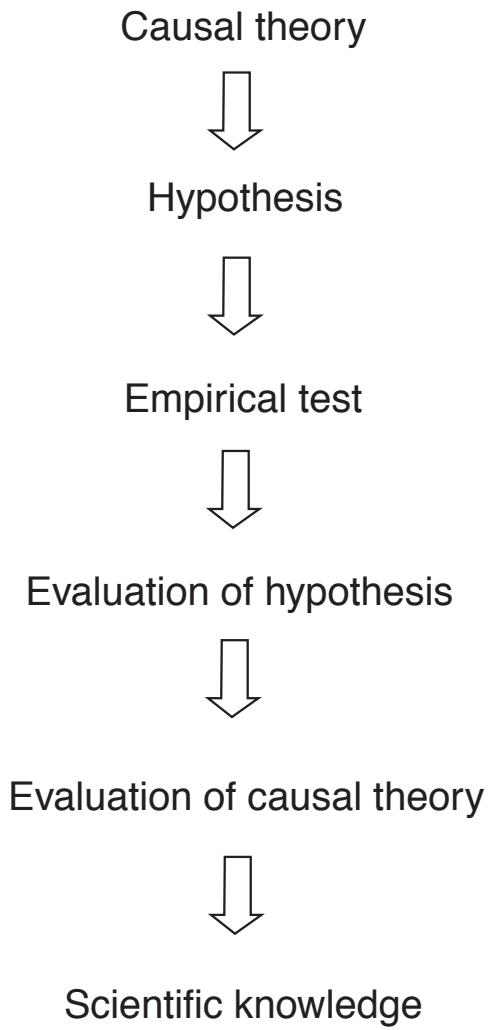
Why should we bother with f?

1. Prediction: $\hat{Y} = \hat{f}(X)$, where $\hat{f}$ is a *black box*.
2. Inference: How Y is changing as a function of X.

- Depending on whether the ultimate goal is prediction, inference or a mix of both, we may deploy different methods for estimating f.
- Also depending on the ultimate goal you may or may not care about evaluating the causal relationship between Y and X.

Research Design

Scientific approach: Searching for causal explanations



Research design

Does X cause Y?

- If we want to test whether X causes Y there are several strategies, or **research designs** that researchers can employ toward that end.
- The goal of all types of research designs is to help us evaluate how well a theory fares as it makes its way over the four causal hurdles—that is, to answer as conclusively as is possible the question about whether X causes Y.

Two broad approaches to designing research

- Experimental design.
- Observational study.

Benchmark of randomized trials

Do hospitals make people healthier?

The National Health Interview Survey (NHIS):

1. During the past 12 months, was the respondent a patient in a hospital overnight?
2. Would you say your health in general is excellent, very good, good, fair, poor?

- Simple reading of the NHIS results suggests that hospitals make people sicker. But people who go to the hospital are probably less healthy to begin with.

Do hospitals make people healthier? Formalized

- Hospital treatment is described by a binary random variable $D_i = 0, 1$
- Health status (outcome) is Y_i
- Is Y_i affected by D_i?

Potential outcomes framework

- In idealized world we imagine what might have happened to a person who went to the hospital if he/she had not gone, and vice versa.

Rubin (1974):

Intuitively, the causal effect of one treatment, E, over another, C, for a particular unit and an interval of time from t_1 to t_2 is the difference between what would have happened at time t_2 if the unit had been exposed to E initiated at t_1 and what would have happened at t_2 if the unit had been exposed to C initiated at t_1: 'If an hour ago I had taken two aspirins instead of just a glass of water, my headache would now be gone,' or 'because an hour ago I took two aspirins instead of just a glass of water, my headache is now gone.' Our definition of the causal effect of the E versus C treatment will reflect this intuitive meaning.

Things to note:

- Potential outcomes and covariates are fixed for each i. Treatments and response indicators are stochastic.
- Effects are defined by letting only treatments vary, not units.
- Thus, causal effects are defined only for units that can conceivably receive different treatment values.
- The test for the above is "manipulation" (Holland,1986).

Formalizing potential outcomes framework

For any individual there are two potential health outcomes:

$$\text{Potential outcome} = \begin{cases} Y_{1i} & \text{if } D_i = 1 \\ Y_{0i} & \text{if } D_i = 0 \end{cases}$$

The difference between two potential outcomes Y_{1i} and Y_{0i} is the causal effect of going to the hospital for individual i.

Observed outcome as a combination of potential outcomes

$$Y_i = \begin{cases} Y_{1i} & \text{if } D_i = 1 \\ Y_{0i} & \text{if } D_i = 0 \end{cases} = Y_{0i} + (Y_{1i} - Y_{0i})D_i \qquad (1)$$

- where $(Y_{1i} - Y_{0i})$ is the causal effect of hospitalization for an individual.
- Since we never observe both potential outcomes for any one person, we must compare the average health effects for two groups (hospitalized and not).

Fundamental problem of causal inference (Holland, 1986):

For each i potential outcomes for all treatments exist, but we only observe the potential outcome for the treatment value that i receives.

- "Scientific solution": Use theory to determine when units are interchangeable and create comparisons.
- "Statistical solution": Study averages.

The comparison of average health conditional on hospitalization status:

$$\underbrace{E[Y_i|D_i=1]-E[Y_i|D_i=0]}_{\text{Observed difference in average health}} = \underbrace{E[Y_{1i}|D_i=1]-E[Y_{0i}|D_i=1]}_{\text{Average treatment effect on the treated (ATT)}} \underbrace{+E[Y_{0i}|D_i=1]-E[Y_{0i}|D_i=0]}_{\text{Selection bias}}.$$

where $E[Y_{1i}|D_i=1]-E[Y_{0i}|D_i=1]=E[Y_{1i}-Y_{0i}|D_i=1]$ is the average causal effect of hospitalization on those who were hospitalized.

Selection bias and random assignment

Selection problem is solved via random assignment because it makes D_i independent of potential outcomes.

$$\begin{aligned} E[Y_i|D_i=1] - E[Y_i|D_i=0] &= E[Y_{1i}|D_i=1] - E[Y_{0i}|D_i=0] \\ &= E[Y_{1i}|D_i=1] - E[Y_{0i}|D_i=1] \\ &= E[Y_{1i} - Y_{0i}|D_i=1] \\ &= E[Y_{1i} - Y_{0i}]. \end{aligned}$$

by virtue of independence of Y_{0i} and D_i we swap $E[Y_{0i}|D_i=1]$ for $E[Y_{0i}|D_i=0]$ in the second line.

Solving the most important problem of empirical research

Random assignment of D_i eliminates selection bias

Regression analysis of experiments

Assume that the treatment effect is constant (the same for all individuals), $y_{1i} - y_{0i} = \rho$. We can rewrite (1) as

$$Y_i = \underbrace{\alpha}_{E[Y_{0i}]} + \underbrace{\rho}_{(Y_{1i}-Y_{0i})} D_i + \underbrace{\eta_i}_{Y_{0i}-E(Y_{0i})}, \quad (2)$$

where η_i is the random part of Y_{0i}.

Evaluating conditional expectation under $D_i = 0, 1$

$$E[Y_i|D_i = 1] = \alpha + \rho + E[\eta_i|D_i = 1]$$
$$E[Y_i|D_i = 0] = \alpha + E[\eta_i|D_i = 0],$$

so that

$$E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = \underbrace{\rho}_{\text{Treatment effect}} + \underbrace{E[\eta_i|D_i = 1] - E[\eta_i|D_i = 0]}_{\text{Selection bias}}.$$

- Selection bias is the correlation between the regression error, η_i, and the regressor, D_i.
- Since,

$$E[\eta_i|D_i = 1] - E[\eta_i|D_i = 0] = E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0],$$

this correlation reflects the difference in (no-treatment) potential outcomes between those who get treated and those who don't.

Some drawbacks of experimental design

1. Can we really randomly assign treatment (X) to subjects?
2. What about *external validity*?
   - Samples of convenience and replication
   - External validity of the stimulus
3. Are there ethical considerations?

Observational studies

- If we cannot evaluate causal theories in a controlled setting like an experiment, we have to take the world as it already is, and use what are called *observational studies*.
- An observational study is a research design in which the researcher does *not* have control over values of the independent variable, which occur naturally. However, it is necessary that there be some degree of variability on the independent variable between cases, as well as variation in the dependent variable.
- Some maintain that, in the absence of experiments, we cannot demonstrate causality with any degree of confidence, but only correlation. Many researchers think that this is a bit too strong of a statement.

What makes a good research project?

1. What is the causal relationship of interest?
2. What is the experiment that could ideally be used to capture the causal effect of interest?
3. What is your identification strategy?
4. What is your mode of statistical inference?

What is the causal relationship of interest?

- The link between hospitalization and health outcomes;
- The link between education and wages;
- The link between institutions and growth.

What is the experiment that could ideally be used to capture the causal effect of interest?

- Randomly assigning incentives to finish school (go to college);
- Randomly assigning colonial institutions;
- Randomly assigning leadership qualities;
- Largely hypothetical, but help identify causal mechanism in play
- If you cannot answer your research question with an idealized, hypothetical experiment, you are unlikely to do it with observational data!
- FUQs: fundamentally unidentified questions (e.g. the effect of start age on first grade scores, the effect of hospital admission on health).

What is your identification strategy?

Identification strategy:

The manner in which a researcher uses observational data (i.e., data not generated by a randomized trial) to approximate a real experiment.

What is your mode of statistical inference?

- Population to be studied.
- Sample to be used.
- Assumptions made when constructing estimators and standard errors.

Review of basic of statistical theory

The probability framework for statistical inference

- Population, random variable, and distribution
- Moments of a distribution (mean, variance, standard deviation, covariance, correlation)
- Conditional distributions and conditional means
- Distribution of a sample of data drawn randomly from a population

Population, random variable, and distribution

- Population
  - The group or collection of all possible entities of interest (school districts)
  - We will think of populations as infinitely large
- Random variable Y
  - Numerical summary of a random outcome (district average test score, district STR)

Population distribution of Y

- The probabilities of different values of Y that occur in the population, e.g. $Pr[Y = 650]$ (when Y is discrete)
- or: The probabilities of sets of these values, e.g. $Pr[640 \leq Y \leq 660]$ (when Y is continuous).

Moments of a population distribution: mean, variance, standard deviation, covariance, correlation

- mean = expected value (expectation) of $Y = E(Y) =$ long-run average value of Y over repeated realizations of Y
- variance = $E(Y - E[Y]) =$ measure of the squared spread of the distribution
- standard deviation = $\sqrt{variance}$
- skewness = measure of asymmetry of a distribution: skewness = 0 – distribution is symmetric; otherwise the distribution has long right or left tail.
- kurtosis=measure of mass in tails=measure of probability of large values: kurtosis =3 – normal distribution; greater than 3 – heavy tails ("leptokurtotic")

Two random variables: joint distributions and covariance

- Random variables X and Z have a joint distribution
- The covariance between X and Z is

$$cov(X, Z) = E[(X - \mu_X)(Z - \mu_Z)]$$

- The covariance is a measure of the linear association between X and Z
- $cov(X, Z) > 0$ means a positive relation between X and Z
- If X and Z are independently distributed, then $cov(X, Z) = 0$ (but not vice versa!)

The correlation coefficient is defined in terms of the covariance:

$$corr(X, Z) = \frac{cov(X, Z)}{\sqrt{var(X)var(Z)}} = \frac{\sigma_{XZ}}{\sigma_X \sigma_Z}$$

- $-1 \leq corr(X, Z) \leq 1$
- corr(X,Z) = 1 mean perfect positive linear association
- corr(X,Z) = -1 means perfect negative linear association
- corr(X,Z) = 0 means no linear association

Conditional distributions and conditional means

- Conditional distributions: The distribution of Y, given value(s) of some other random variable, X. E.g. the distribution of test scores, given that STR < 20
- Conditional expectations and conditional moments: conditional mean = mean of conditional distribution = $E(Y|X = x)$ (important concept and notation)
- The difference in means is the difference between the means of two conditional distributions.

Distribution of a sample of data drawn randomly from a population

- We will assume simple random sampling: Choose and individual (district, entity) at random from the population.
- Randomness and data: Prior to sample selection, the value of Y is random because the individual selected is random. Once the individual is selected and the value of Y is observed, then Y is just a number – not random.
- Because individuals no. 1 and no. 2 are selected at random, the value Y for the first has no information content for the second. Thus, Y_1 and Y_2 are independently distributed, and if they come from the same distribution then they are also identically distributed.
- That is, under simple random sampling, Y_1 and Y_2 are independently and identically distributed (i.i.d.).

Estimation

- $\bar{Y}$ is the natural estimator of the mean.
- $\bar{Y}$ is a random variable, and its properties are determined by the sampling distribution of $\bar{Y}$.
- The distribution of $\bar{Y}$ over different possible samples of size n is called the sampling distribution of $\bar{Y}$.
- The mean and variance of $\bar{Y}$ are the mean and variance of its sampling distribution, $E(\bar{Y})$ and $var(\bar{Y})$.
- The concept of the sampling distribution underpins much of statistical testing.

The sampling distribution of $\bar{Y}$ when n is large

For small sample sizes, the distribution of $\bar{Y}$ is complicated, but if n is large, the sampling distribution is simple!

1. As n increases, the distribution of $\bar{Y}$ becomes more tightly centered around μ_Y (the Law of Large Numbers)
2. Moreover, the distribution of $\bar{Y} - \mu_Y$ becomes normal (the Central Limit Theorem)

A gentle introduction to R and RMarkdown